AD-A185 837

(12)

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION U	1b. RESTRICTIVE MARKINGS NA
2a. SECURITY CLASSIFICATION AUTH NA	3. DISTRIBUTION/AVAILABILITY OF REPORT Distribution Unlimited
2b. DECLASSIFICATION/DOWNGRADI[illegible]CHEDULE	
4. PERFORMING ORGANIZATION REP[illegible]UMBER(S) NA	5. MONITORING ORGANIZATION REPORT NUMBER(S) NA

6a. NAME OF PERFORMING ORGANIZATION California Institute of Technology	6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION Office of Naval Research
6c. ADDRESS (City, State, and ZIP Code) Division of Chemistry 147-75 California Institute of Technology Pasadena, California 91125		7b. ADDRESS (City, State, and ZIP Code) 800 North Quincy Street Arlington, Virginia 22217-5000
8a. NAME OF FUNDING/SPONSORING ORGANIZATION Office of Naval Research	8b. OFFICE SYMBOL (If applicable) ONR	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER N00014-86-K-0755
8c. ADDRESS (City, State, and ZIP Code) 800 N. Quincy Street Arlington, Virginia 22217-5000		10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. / PROJECT NO. / TASK NO. / WORK UNIT ACCESSION NO.

11. TITLE (Include Security Classification)

Biopolymers: Proteins and Nucleic Acids

12. PERSONAL AUTHOR(S)

John H. Richards, J. N. Abelson, L. E. Hood, M. I. Simon, J. L. Campbell, P. B. Dervan and W. A. Goddard

13a. TYPE OF REPORT Annual	13b. TIME COVERED FROM 9/15/86 TO 9/14/87	14. DATE OF REPORT (Year, Month, Day) September 15, 1987	15. PAGE COUNT 27 (+ reprints)

16. SUPPLEMENTARY NOTATION

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	DNA Binding; Synthetic Peptides; Reagents for DNA Cleavage; Nucleic Acids; Protein Structure; Deoxyribonucleic acids

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

The work focuses on learning the principles that govern interactions between proteins and nucleic acids. With these principles as guides we are synthesizing peptides (of about 50 amino acids) that bind to specific regions of DNA. Various reactive functionalities are being attached to the synthetic peptides to generate reagents that cleave DNA specifically at the site to which the peptide binds.

The work also involves biophysical and theoretical studies of the protein/nucleic acid complexes in order to expand our understanding of the principles of protein binding to nucleic acid as well as the development of improved procedures for the chemical synthesis of peptides. Keywords:

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION U
22a. NAME OF RESPONSIBLE INDIVIDUAL Dr. Robert W. Newburgh	22b. TELEPHONE (Include Area Code) (202) 696-4986 / 22c. OFFICE SYMBOL ONR

DD FORM 1473, 84 MAR 83 APR edition may be used until exhausted. All other editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE

87 9 2 U [illegible] 3

ANNUAL REPORT

September 15, 1987

Biopolymers: Proteins and Nucleic Acids

California Institute of Technology
Pasadena, California 91125

Co-Investigators: John N. Abelson, Leroy E. Hood and Melvin I. Simon, Division of Biology; Judith L. Campbell, Divisions of Biology and Chemistry; Peter B. Dervan, William A. Goddard and John H. Richards, Division of Chemistry

Contract: N00014-86-K-0755

Summary

The past year has seen considerable progress in our understanding of the factors that influence interactions between proteins and nucleic acids together with our ability to design and synthesize peptides and their derivatives that bind to specific DNA sequences and cleave the DNA at these sites. Such highly specific reagents will find important applications in many areas of modern molecular biology including such monumental undertakings as the sequencing of the entire human genome which is the focus of so much present attention.

One such system which we are studying intensively is Hin recombinase, an enzyme that acts *in vivo* to invert a 996 bp region of DNA and in doing so to turn on or off expression of the bacterial flagellin H2 gene. During this year we have been able to come to a number of important conclusions about the mode of action of this enzyme and which region of it (the C-terminal 52 amino acids) is especially involved in DNA binding and recognition (Simon). A 52 mer peptide constituting this C-terminal region of Hin recombinase has been synthesized and shown to bind with the specificity of the intact enzyme. Moreover, to it have been added various reactive groups that allow the synthetic, derivatized peptide to act as a DNA cleavage reagent (Dervan). We have also extended this work to other related peptides and have considerably expanded the types of functionalities that have been attached to broaden the range of available DNA cleaving reagents.

To understand the interactions that form the basis of our ability to design these DNA binding/cleaving molecules, extensive structural studies are underway involving theoretical and magnetic resonance approaches. Briefly, the 2D nmr spectra of an oligonucleotide duplex that mimics the DNA binding site of the Hin recombinase peptide shows that significant distortions from the normal B helix exist in the middle region of the duplex suggesting that some structural feature of the binding site preexists in the DNA binding site and plays an important role in the binding of the Hin recombinase (Richards). The theoretical structural work has developed new computing approaches to predictions of secondary and tertiary structures based on stochastic (Monte Carlo) and deterministic (molecular dynamic) methods to search for global energy minima (Goddard). Computer modeling has also been used to correlate 2D nmr spectra from putative 3D structures of DNA and proteins.

Proteins that bind to single stranded DNA and to RNA form a second important aspect of the overall program. In this area considerable progress has been

made on single stranded DNA binding proteins (Campbell) and on proteins (such as tRNA ligase) that are responsible for tRNA splicing in eukaryotes (Abelson).

Lastly, the preparation of the synthetic peptides that form the central core of much of the work requires that we have readily available at Caltech the most sophisticated synthetic methodologies. To this end we have developed in the past year improved protocols for the steps necessary to add amino acid residues to a growing peptide chain. These improvements have considerably enhanced the stepwise yield and the purity of the resulting synthetic peptides (Hood).

Melvin I. Simon

During the past year we have made a great deal of progress in two areas: (1) defining the nature of the region of the DNA that binds the peptide derived from Hin recombinase and (2) in modifying the peptide and engineering it so that it has both good binding activity and contains a DNA cutting domain. These experiments clearly indicate that the approach that we are taking is the correct one. It should be possible for us to use modern techniques of peptide synthesis, molecular genetics and organic chemistry to develop new molecules with novel and useful features. Our long range goal is to convert the Hin recombinase peptide into an enzyme function which will recognize long stretches of DNA sequence, specifically and introduce cleavages in the DNA adjacent to those long sequences. This accomplishment would demonstrate both our ability to engineer peptides for specific functions and would provide a new kind of restriction enzyme that would be extremely useful in human gene mapping. Finally, in the context of the manipulation of the Hin peptide we will learn a great deal about the mechanisms involved in protein-DNA recognition.

A. Characterization of Hin Recombinase Binding

The Hin-recombinase acts *in vivo* to invert a 996 bp region of DNA and to turn on or to turn off the expression of the bacterial flagellin H2 gene. It does this by mediating site-specific recombination between two hix sites that flank the invertible segment and DNA inversion results from the recombination reaction. The Hin recombinase binds to each 26 base pair *hix* site. The nucleotide sequences within each site have an axis of symmetry with the recombinase molecules binding to each half site. The binding occurs through the C terminal 52 amino acids of the Hin recombinase. Two kinds of approaches were taken to more carefully define the nature of the binding site; one was to study binding *in vitro* by using a variety of chemical techniques. These included measurements of the effects of the bound peptide in interfering with chemical methylation, ethylation and further modification of specific nucleotide base pairs, as well as measurement of protection against nuclease cleavage adjacent to specific nucleotide base pairs as a result of the binding of the 52 amino acid peptide or the Hin recombinase to the *hix* sites. As a result of these studies we came to a number of conclusions:

1. Hin recombinase binds in a cooperative manner to both halves of the symmetrical site. When only one half site is present, binding affinity is lower by a factor of 100 than when both sites are present. The 52 amino acid peptide corresponding to the C-terminus of Hin-recombinase binds independently to each of the two half sites.

2. Two adenine-thymine (A-T) base pairs in the minor groove of the *hix* half sites are important for binding both of the peptide and of the Hin recombinase. Furthermore, there is strong interaction between the polypeptide and one or

two base pairs that are found in the major groove. There is, therefore, evidence for both major and minor groove interactions.

3. All of the binding data and gel retardation data suggest that aside from cooperative effects, the peptide that contains the C-terminal 52 amino acids of the Hin recombinase shows the same binding properties as the complete enzyme. Thus, it appears that this 52 amino acid peptide is responsible for the nucleotide sequence specific recognition required for recombinase binding.

While the chemical studies have defined some of the sites on the DNA that interact with the polypeptide *in vitro*, we wanted to determine if these same relationships existed in the cell *in vivo*. In order to do this, we developed a series of assays that allow us to measure the affinity of Hin binding to a variety of mutant *Hix* sites. Over 120 individual mutant· which include specific deletions, insertions and base pair changes have been sequenced and characterized. All of the *in vivo* data is consistent with the *in vitro* studies; that is, insertion of even one or two base pairs between the symmetrical half sites markedly decreases the binding of the Hin recombinase. These results suggest that the position of the binding sites on the DNA relative to each other is very important in enhancing cooperative interactions between Hin protein bound at the symmetrical site. Furthermore, base substitution mutations have demonstrated that changes of the minor groove AT base pairs to GC base pairs markedly affect binding. Relatively strong effects of a change of the base pairs that were previously found to be protected by Hin binding in the major groove were also found as a result of mutagenesis. All of the mutagenesis results are in clear agreement with the findings *in vitro* by the chemical techniques. We have further characterized an additional in recombinase binding site. This site is found adjacent to the gene that encodes the Hin protein and it has been suggested that the binding of Hin recombinase to this site may act to regulate the synthesis of Hin.

This work is currently being prepared for publication:

1. Hughes, K., Youderian, P. and Simon, M. I. Phase variation in Salmonella: analysis of Hin recombinase and *hix* recombination site interaction *in vivo*. Manuscript in preparation.

2. Glasgow, A. and Simon, M. I. Characterization of the recombinase DNA interactions *in vitro*. Manuscript in preparation.

B. Characterization of the Hin Recombinase Peptide which Binds DNA

In order to characterize the domain of the Hin recombinase that binds DNA, we synthesized a peptide including the C terminal 52 amino acids of Hin. This peptide was shown to bind the Hix site and to give the same pattern of chemical protection as the whole Hin-recombinase with a binding constant about 5% lower than that of Hin recombinase binding to a single Hix half site. We also synthesized smaller versions of this peptide and found that deletion of two amino acids from the N-terminus strongly affected the ability of the peptide to bind to DNA. These two amino acids do not fall within the helix-turn-helix structure which has been implicated in DNA binding in other systems. Thus, these changes suggest that there is yet another element in the 52 amino acid peptide that is required in order to bind to DNA.

To further develop our ability to engineer this peptide in order to make it useful as a DNA cleaving reagent, we collaborated with Dr. Peter Dervan and with Suzanna Horvath. Dr. James Sluka in Peter Dervan's laboratory synthesized an

EDTA derivative at the N-terminus of the 52 amino acid peptide. Thus, he was able to introduce an Iorn atom specifically at the N-terminal of the peptide. When this peptide Iorn complex was mixed with DNA and the proper reducing agents added we found that the DNA was specifically cleaved adjacent to the binding sites of the Hin peptide. These results allow us to develop a model for a mechanism by which Hin binds DNA. Furthermore, they provide us with a first generation DNA-cleaving peptide and suggest new possibilities for engineering the peptide.

This work has been submitted for publication:

1. Bruist, Horvath, S., Hood, L. E., Steitz, T. and Simon, M. I. (1987) *Science* **235**, 777.

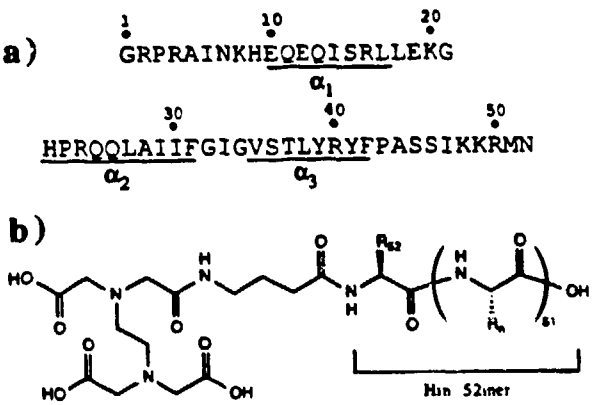
2. Sluka, J., Horvath, S., Bruist, M. F., Simon, M. I. and Dervan, P. B. (1987) *Science*, submitted.

In ongoing experiments, we are trying to modify the Hin peptide so as to increase its affinity for DNA. This is being done in a number of ways; both by chemical techniques and by mutagenetic techniques *in vivo* and *in vitro*. We are also trying to carefully characterize the nature of the interaction of the peptide with the DNA. This is being done in a variety of ways. One of which is an attempt to crystallize the peptide together with DNA. Preliminary results suggest that small crystals can be obtained and we are pursuing this approach intensely.

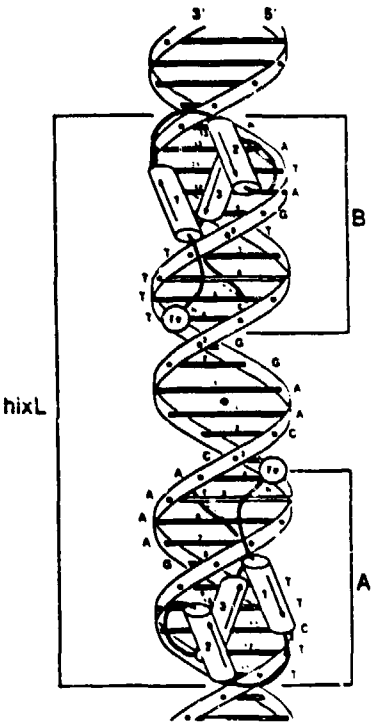
Peter B. Dervan

1. Synthesis of a Sequence Specific DNA Cleaving Peptide

J.P. Sluka, S.J. Horvath, M.F. Bruist, M.I. Simon, P.B. Dervan, *Science* in press. A synthetic 52 residue peptide based on the sequence specific DNA binding domain of Hin recombinase (139-190) has been equipped with ethylenediaminetetraacetic acid at the N-terminus. In the presence of Fe(II), this synthetic EDTA-peptide cleaves DNA at Hin recombination sites. The cleavage data reveals that the N-terminus of the 52mer is bound in the *minor* groove of DNA near the symmetry axis of Hin recombination sites. This work demonstrates the construction of a hybrid peptide combining two functional domains: sequence specific DNA binding and DNA cleavage.



EG52



1.5μM EG52-Fe

5' TTTATTGGTTCTTGAAAACCAAGGTTTTTGATAAAGCAATCCTCCATGAGAA
3' AAATAACCAAGAACTTTTGGTTCCAAAAACTATTTCGTTAGGAGGTACTCTT

2. Peptide:DNA Recognition. Mapping the DNA Binding Domain of Hin Recombinase Using the Affinity Cleaving Method

J.P. Sluka, S.J. Horvath, M.I. Simon, P.B. Dervan manuscript in preparation (*Cell*). Twelve synthetic peptides (49,50,51,52,56 and 60 residues in length) have been synthesized corresponding to the DNA binding domain of Hin recombinase (residues 141-190, 140-190, 139-190, 138-190, 134-190, 130-190) with and without EDTA attached at the N-terminus. From MPE·Fe(II) footprinting of six peptides and affinity cleaving of six EDTA-peptides we are determining the significance of the Arg-Pro-Arg-Gly (residues 49,50,51,52) for sequence specific binding of the N-terminus of the DNA binding domain in the minor groove of DNA. We have made the interesting observation that the N-terminus Gly can be removed without consequence. However, the removal of the next Arg appears to decrease binding by at least a factor of 10. Moreover, peptides 56 and 60 residues in length appear to have *movable* arms revealed by cleavage patterns on sequencing gels. This is our first insight on the "linker" region between the binding domain and cleavage domain of Hin.

3. Design of a Protein that Binds and Cleaves DNA Sequence Specifically

D. Mack and P.B. Dervan, in preparation (*Nature*). A synthetic peptide 55 residues in length constructed of wholly natural α-amino acids combines the DNA binding domain of Hin recombinase (139-190) and the copper binding site of serum albumin, Gly-Gly-His. This designed protein, Gly-Gly-His (Hin 138-190) binds the Hin recombinase half site (13 bp in size) and in the presence of 1 equiv of Cu(II), hydrogen peroxide/ascorbate cleaves the DNA at those sites. The major site of cleavage suggests that a *non-diffusible* oxidizing species in the minor groove of DNA near the symmetry axis of Hin recombination site is responsible for the cleavage.

4. Peptide: DNA Recognition: N-Terminus Location of γδ Resolvase (141-183) on DNA

K. Graham and P.B. Dervan, in preparation (*JACS*). A synthetic 42 residue peptide based on the DNA binding domain of the site specific recombination protein γδ resolvase (141-183) has been equipped with EDTA at the N-terminus. In the presence of Fe(II), this synthetic EDTA-peptide cleaves DNA at the γδ resolvase recombination sites (TGTCYNNTA). The cleavage data reveals that the N-terminus of the 42 mer is bound in the minor groove of DNA near the symmetry axis side of γδ resolvase recombination sites.

A. Specific Aims: October 1987-88

We have *three* major goals for the next year:

1. Peptide: DNA Recognition

Based on the success of the synthetic methodology for attaching EDTA to synthetic peptides, we will use the affinity cleaving method to understand the relative contributions of several key amino acids and structural domains (such as helix-turn-helix) for Hin (138-190) and γδ resolvase (141-183). This will be accomplished by the synthesis of ~ 20 different peptides with judiciously chosen single amino acid substitutions and helix switch experiments between the two peptides.

2. Protein Engineering

We will construct hybrid synthetic peptides that combine DNA *binding domains* from larger proteins with *cleavage domains* from other proteins to form substances not found in nature, yet consisting of wholly naturally occurring α-amino acids.

3. Metalloregulation of Peptide Binding on DNA

We will construct synthetic hybrid proteins that bind DNA in a sequence specific fashion only in the presence of certain metal cations as allosteric effectors.

John H. Richards

Recently a synthetic 52 residue peptide identical to the C-terminal segment of Hin recombinase has been shown to bind to Hin recombination (HixL) sites and to inhibit Hin activity (1, 2). Several smaller DNA sequences (no bigger than 14 bp) also specifically form binding complexes with the 52-mer peptide. To elucidate the structural features of these DNA-protein complexes forms the central objective of this proposal; such knowledge would add significantly to our understanding of general aspects of DNA-protein interactions.

For these studies we plan to use high resolution nmr as the primary method to study the binding complex described above which is of an appropriate size for this approach. Moreover, nmr has the distinct advantage of allowing one to observe biologically active substrates in real time under actual physiological conditions.

A. DNA

```
                    (5)————————————————
HixL  TTATT(1)GGTTCTTGAAAACCAA(3)GGTTTTTGATAAAGCAATC

                    (6)————————————————
      AATAA(2)CCAAGAACTTTTGGTT(4)CCAAAAACTATTTCGTTAG
```

The DNA sequences selected in this research are 1, 2, 3, 4, 5, 6 as shown above; among them duplex 1.2 and 3.4 cover the sequences of HixL protected from DNase I cleavage by the 52-mer peptide (**bold letters**). 5.6 covers the strongest binding region in a footprinting experiment that showed a 22 bp region protected by the 52-mer peptide.

The DNA's used in the preliminary studies were synthesized using the solid-phase phosphoramidite method, carried out on a 10 micromole scale. Some of the difficulties in purifications are the separation of the DNA from the minor nondeprotected impurities. This problem was solved by prolonged deprotection in saturated ammonia, use of reverse phase HPLC and passage through an ion-exchange oligonucleotide column. This protocol gives excellent resolution of very pure samples and easily allows preparation of 300-600 $O.D._{260}$ units (10-20 mg) of single-strand DNA 1-6.

1. Double Stranded DNA

As well as the planned nmr structural studies of the unsymmetrical duplexes, we have completed the analysis of a self-complementary oligomer 7: GGTTTTCGAAAACC, which is analogous to 1.2 with just one base pair shift.

In ion exchange HPLC, 7 shows a dramatically different retention time compared to the oligomers that are not self-complementary, indicating that 7 exists as a duplex in solution, in accordance with previous studies on interconversions of hairpin and duplex DNA conformers (3).

2. DNA Structure in Solution by 2D NOESY

Figure 1 shows a NOESY spectrum, obtained at the Southern California Regional NMR Facility, recorded at 298 K in D_2O with a mixing time of 200 ms. All non-labile proton resonances and cross peaks can be assigned except for a few 5'5"H's and 2H's of adenosine. Typical patterns in sugar 1'H-base 6/8H (Figure 5) and 2'2"H-6/8H (Figure 6) regions show that this DNA exists as B-type duplex under these conditions.

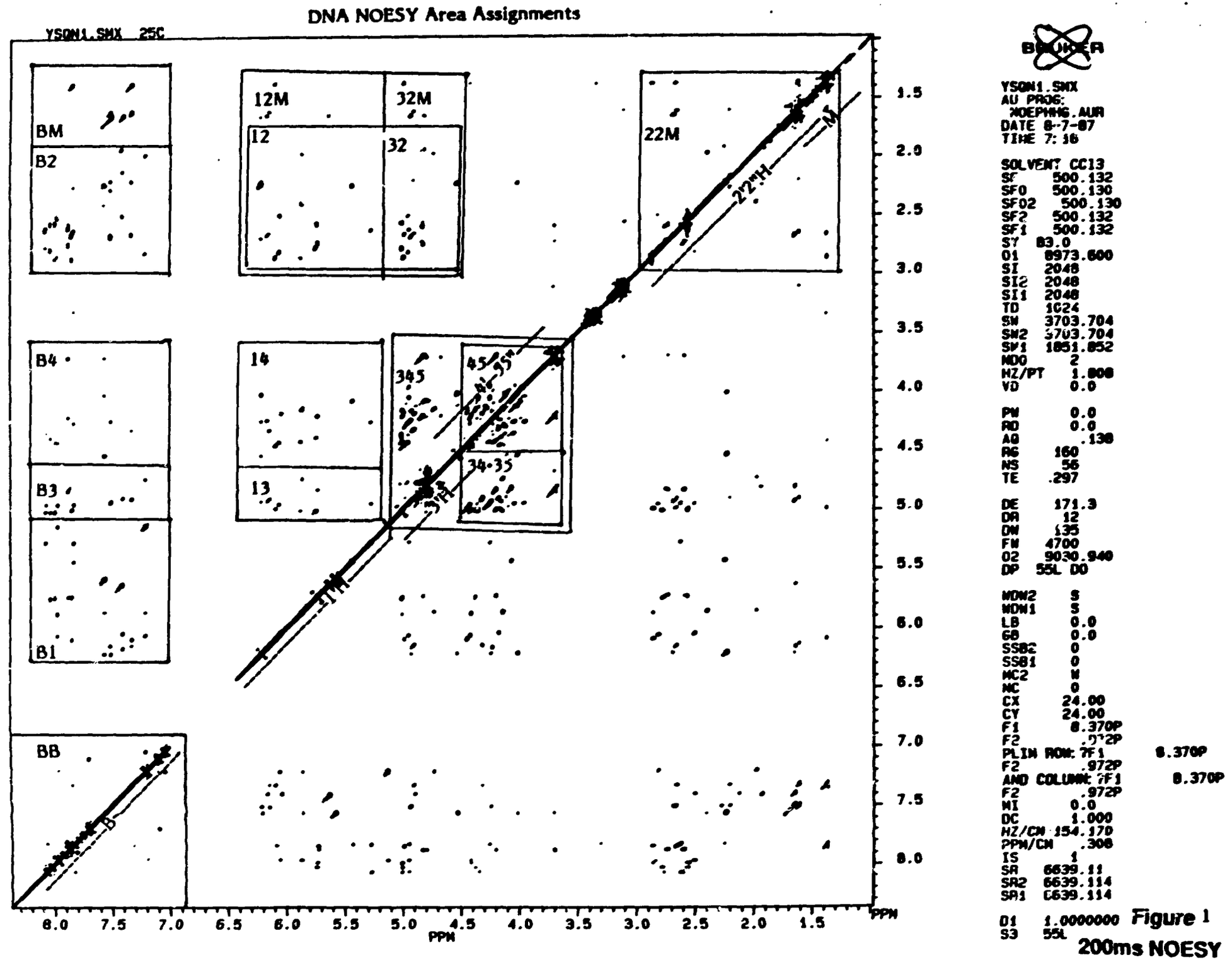
With a mixing time of 200 ms the NOE spectra contain a certain amount of spin diffusion and cannot therefore be used directly to determine distances. Accordingly, another NOESY was determined under the same solution conditions, but with a mixing time of 50 ms; this excludes spin-diffusion in the NOE build-up, so that the observed NOE's are directly proportional to the reverse sixth power of distances (Figure 2). Our estimate is that the cut-off distance (where NOE = 0) is around 3.5 Å for these spectra.

The calculated sequential internucleotide distance between sugar 2H' of the nth base and base proton 6/8H of the (n + 1)th base is 2.1-2.2 Å in the standard B-type DNA (4), leading to a significant NOE between protons in these relative positions. Indeed, most of the observed NOE's fall within the range expected for such distances. However, for protons on bases 5 and 6 and bases 6 and 7, the small observed NOE's (Figure 2) imply interbase distances near 3.5 Å suggesting some distortion from a normal B structure in the middle region of the duplex. Thus some structural feature of the duplex, preexisting at the HixL site, may play an important role in the specific binding of the Hin protein.

B. Peptides

Hin 52-mer peptide was synthesized chemically on an ABI Peptide Synthesizer and purified using reverse phase HPLC on a C-4 column. A 1-D nmr study as a function of temperature of the peptide thus purified suggests that it exists largely as a random coil in solution, in the absence of DNA, a result which agrees with CD measurements of the peptide under similar conditions. Taken together with the dramatic specificity with which the Hin peptide recognizes the HixL DNA sequence, these nmr and CD results suggest that interactions between peptide and DNA are essential to induce folding of the peptide into a three-dimensional structure that binds specifically to DNA. (Putatively, this structure contains the helix-turn-helix motif of many DNA binding proteins.)

Even after HPLC purification there is a significant possibility that the peptide contains impurities that mask the nmr spectrum of a folded form of the peptide. Therefore, to separate peptides that bind to the HixL site from contaminants, we will



BRUKER

YSON1.SMX
AU PROG:
NOEPHHG.AUR
DATE 8-7-87
TIME 7:16

SOLVENT CCl3
SF 500.132
SFO 500.130
SFO2 500.130
SF2 500.132
SF1 500.132
ST 83.0
O1 8973.600
SI 2048
SI2 2048
SI1 2048
TD 1024
SW 3703.704
SW2 3703.704
SW1 1851.852
NDO 2
HZ/PT 1.808
VD 0.0

PW 0.0
RD 0.0
AQ .138
RG 160
NS 56
TE .297

DE 171.3
DR 12
DW 135
FW 4700
O2 9030.940
DP 55L DO

NDW2 S
NDW1 S
LB 0.0
GB 0.0
SSB2 0
SSB1 0
MC2 M
MC 0
CX 24.00
CY 24.00
F1 8.370P
F2 .972P
PLIN ROW: F1 8.370P
F2 .972P
AND COLUMN: F1 8.370P
F2 .972P
MI 0.0
DC 1.000
HZ/CM 154.170
PPM/CM .308
IS 1
SR 6639.11
SR2 6639.114
SR1 6639.114

D1 1.0000000
S3 55L

Figure 1
200ms NOESY

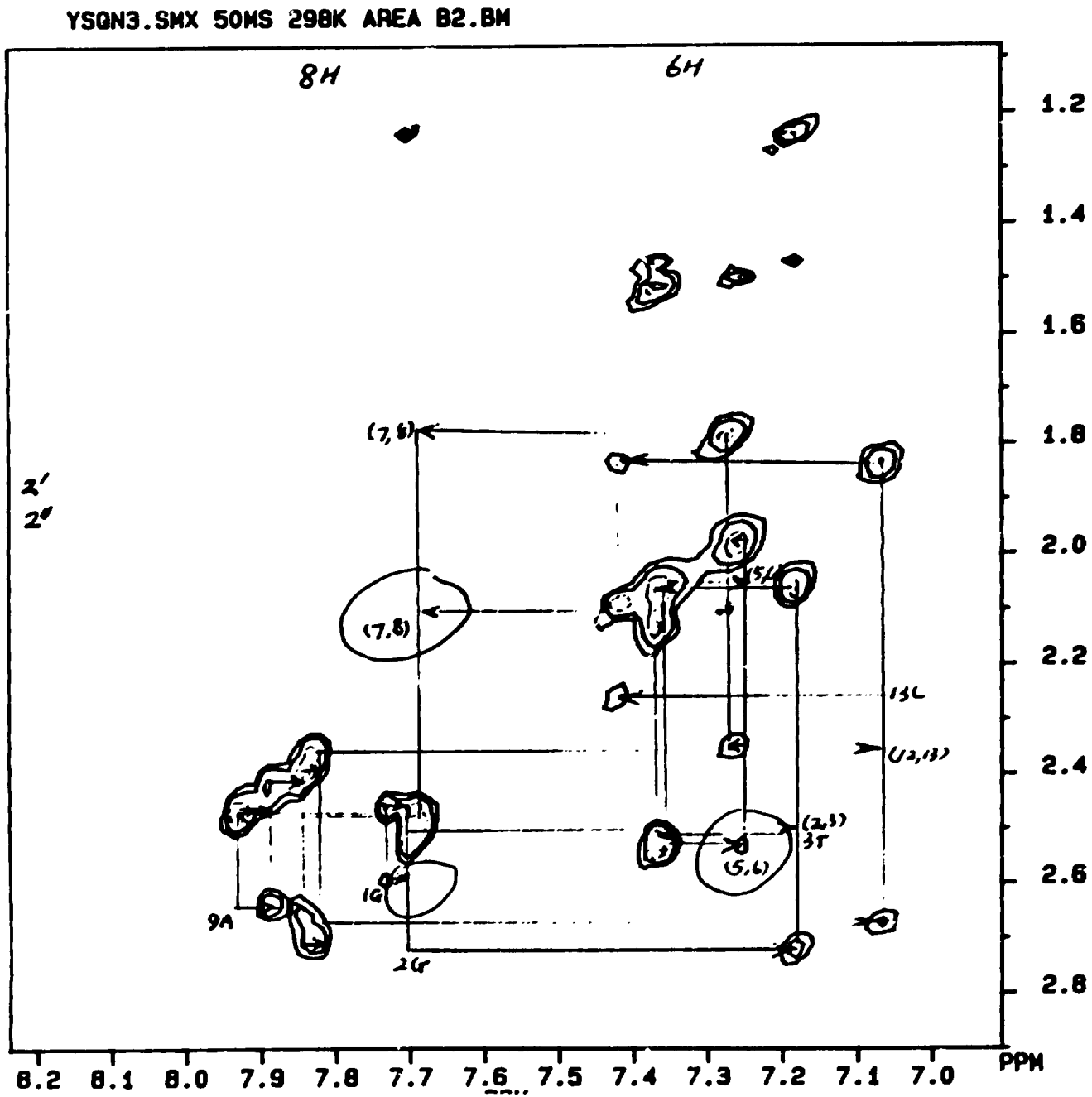


Figure 2 50ms NOESY

use affinity chromatography with a 1.2 DNA duplex (the HixL site) linked to a resin by annealing a duplex with a 5'-penta A extension to a resin with a covalently linked 5'- poly T segment.

```
5'-AAAAA-GGTCCTTGAAAACC

Resin--TTTTT CCAAGAACTTTTGG-3'
```

C. Protein-DNA Complex

To study the solution structure of the DNA peptide complex by ^{1}H 2-D nmr will involve the assignment of nearly 600 non-labile protons (109 from the 14 mer duplex 7 and about 490 from the peptide). To overcome some of the resulting problems in assignments, we plan to use nmr techniques such as a multiquantum filter and an X-filter (5) together with observations of derivatives containing specific ^{13}C, ^{15}N and ^{19}F labels (6). The ability to synthesize chemically both the DNA and peptides used in this work provides a powerful incentive to the use of such labelled species to establish the solution structure of this DNA-peptide complex. ^{31}P nmr will also be used to study the phosphate backbone.

References

1. Bruist, M. F., Horvath, S. J., Hood, L. E., Steitz, T. A. and Simon, M. I. (1987) *Science* 235, 777.

2. Sluka, J. P., Horvath, S. J., Bruist, M. F., Simon, M. I. and Dervan, P. B. (1987) *Science*, in press.

3. Wemmer. D. E., Chou, S. H., Hare, D. R. and Reid, B. R. (1985) *Nucl. Acids Res.* 13, 3755.

4. Wuthrich, K. (1986) *NMR of Proteins and Nucleic Acids*, John Wiley & Sons, p. 215.

5. Worgotter, E., Wagner, G. and Wuthrich, K. (1986) *J. Am. Chem. Soc.* 108, 6162.

6. Senn, H., Otting, G. and Wuthrich, K. (1987) *J. Am. Chem. Soc.* 109, 1090.

William A. Goddard III

A. Tertiary Structure - First Principles Prediction of Protein Folding

The major problem obstructing the design and synthesis of biological systems is the difficulty in predicting the three-dimensional (3D) folding based only on the sequence of amino acids. The development of protein and DNA synthesizers and the development of techniques for modifying genes and inserting them in systems capable of expression and growth have put us at the stage where we could build nearly any biopolymer. The problem is, which one do we want? To design new materials, we *must* be able to predict 3D structure, since this structure defines the biological and chemical properties.

We are addressing this problem of predicting the biologically significant structures of those proteins for which high-resolution x-ray crystal structures are unavailable. We are attacking the general problem in three ways:

1. **Full Optimization.** We use a combination of stochastic methods (Monte Carlo) and deterministic methods (molecular dynamics) to search for the global energy minima of proteins for which no structural information is known beyond the amino acid sequence (primary structure). This completely general approach is the ultimate solution, but may require several years of development.

2. **Conformationally-Constrained Optimization.** A somewhat less ambitious approach is to use the results of low-resolution x-ray studies (which can determine conformation but do not provide atomic-level resolution) to provide constraints for the molecular dynamics. We use a similar combination of Monte Carlo and molecular dynamics to search the conformational space of proteins but with constraints that reduce the time to determine a structure.

3. **NMR Pairwise-Constrained Optimization.** A third approach is to use the pairwise distances from two-dimensional NMR studies in conjunction with molecular dynamics and energy minimization to predict optimal protein structures consistent with the experimental information.

The NMR pairwise constrained optimization will become an important tool in extracting 3D studies from the NMR facility being set up in association with the DARPA effort. This will allow us to obtain 3D structures for molecules in solution.

The conformationally-constrained optimizations are currently being used to obtain a 3D structure for the 52mer of HIN recombinase attached to DNA. We obtain the conformation by (1) fitting a 3D structure to the C_α coordinates available for a similar protein, CRO, (2) mutating CRO to make the HIN 52mer, and then (3) optimizing a 3D structure for the HIN 52mer, followed by (4) modeling of the HIN into the expected recognition site on the DNA, and (5) optimization of the DNA-HIN complex.

In addition to the standard molecular dynamics techniques (evaluation of forces on the atoms and then solving Newton's equations to obtain new coordinates and velocities), we have developed two new stochastic techniques that are proving essential in solving the problem. To get a sense of the magnitude of the potential difficulties, consider just the case of specifying the two dihedral angles for each C_α of the backbone and imagine that we consider six values for each dihedral angle. This leads to $6^{104} = 8.5 \times 10^{80}$ possible sets of angles. Assuming that one could do a million geometries per second, it would still take 2.7×10^{67} years to calculate all the geometries, and this does not even worry about the side chains!

Our approach is to sample this galactic-sized configuration space using statistical (Monte Carlo) procedures. In one approach, *Brownian torsions* we evaluate the usual energies and forces and then add random torques and angular velocity impulses appropriate for a heat bath at temperature T_B in contact with the torsional degrees of freedom. We find that allowing T_B to be, say, 5000°K, while keeping all the other modes at 300°K will make the torsional degrees of freedom molten. As a result, the usual dynamic equations of motion rapidly sample the torsional degrees of freedom. These dynamics are calculated as a function of time, but we periodically take the current structure and energy minimize to find the optimum geometry corresponding to the current step. We then save the 100 best

structures from a long dynamics run and examine them to obtain the optimum structure of the system. This Brownian torsion procedure works very reliably for smaller polypeptides (12 or less), but for large structures (30 or more), incorrect secondary structures can be locked in, preventing the system from refolding to a better conformation.

This has led to a second approach, *torsional-constrained dynamics.* In this procedure we evaluate all forces (including torsion, van der Waals, electrostatics, hydrogen bonds, and conformational constraints) at each time step in terms of Cartesian coordinates on each atom. These forces are then resolved into torques about each dihedral angle and the equations of motion solved to provide increments in the torsion angles and torsion angular velocities. These are used to predict new torsional angles that are used to predict a new 3D structure. With this 3D structure, we reevaluate Cartersian components of force, resolve them onto torsions, and continue iteratively until the process converges. At this point we do the usual molecular dynamics simulations to get a final structure.

The procedure is still in test, but it appears to have the speed and reliability needed for predicting folding of 50mer's.

B. HIN Recombinase

A very special collaboration that resulted directly from the ONR/DARPA project involves HIN-Recombinase, an enzyme that regulates recombination by causing site-specific DNA inversion (see Figure 1). The protein was isolated, purified, and characterized by Simon and co-workers who eventually were able to find a 52-amino acid sequence that plays a specific role in binding to the HIX-L and HIX-R sites of the gene. A number of experiments (DNA methylation, DNase I and MPE footprinting) were carried out by Simon and others. As part of the ONR/DARPA project, the Dervan group has coupled EDTA to a synthetic version of this 52mer and carried out cleavage studies establishing additional footprinting data. These experiments, plus data of which DNA base pairs are conserved at the HIX sites in various species, provided the data needed for us to model the binding site. To begin the modeling, we started with CRO and modified the three alpha helices (in the helix-turn-helix conformation) to take into account the dramatic effect of the extra PRO groups of HIN-52. We were able to find a binding site consistent with all data. This led to initiating several experiments to test the new model. For example, the Dervan group is attaching their EDTA probe to the carboxy terminus (as opposed to the N-terminus in the original experiment). The new model predicts EDTA cleavage at a site far removed from the highly conserved region. In addition, the theory group has expanded upon these modeling studies to use simulation in predicting the optimum folding at the binding site.

C. Unconstrained Simulation

For fully unconstrained predictions of folding, we are initially focusing on organic polymers (such as polymethylene, polypropylene, polyethers) rather than biopolymers. We have developed some general procedures (related to simulated annealing) for predicting conformation. We find that for isolated polymers having 30 or more atoms along the backbone, there is a very distinct phase transition from a random coil at high temperature to a highly-folded globular form at low temperature (see Figure 2). The transition is related to the strength of interatomic interactions versus the chain rigidity. Indeed, using a mean field theory, we can parameterize the globularity of the polymer (radius of gyration) in terms of a single effective

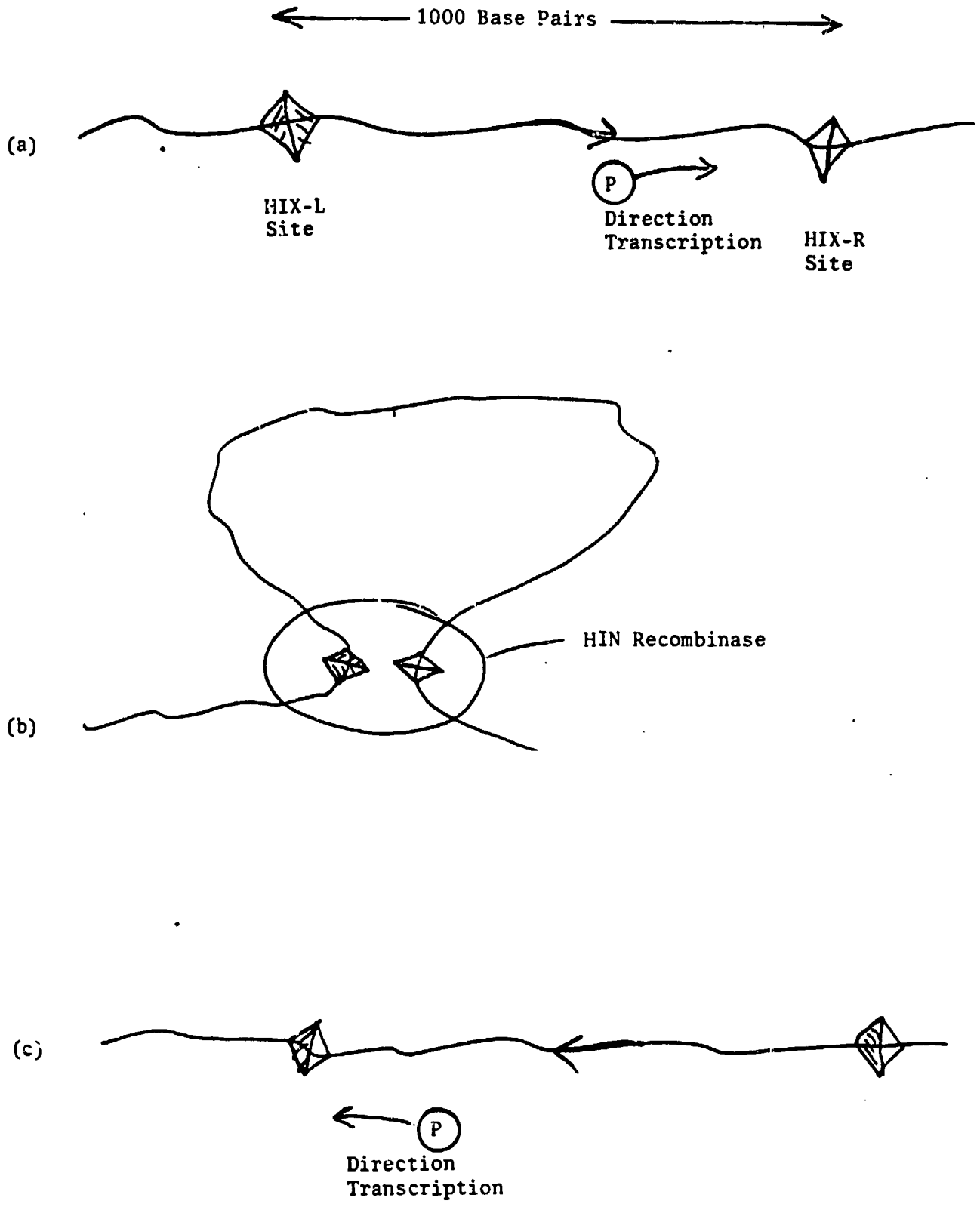


Figure 1. DNA recombination catalyzed by HIN recombinase.

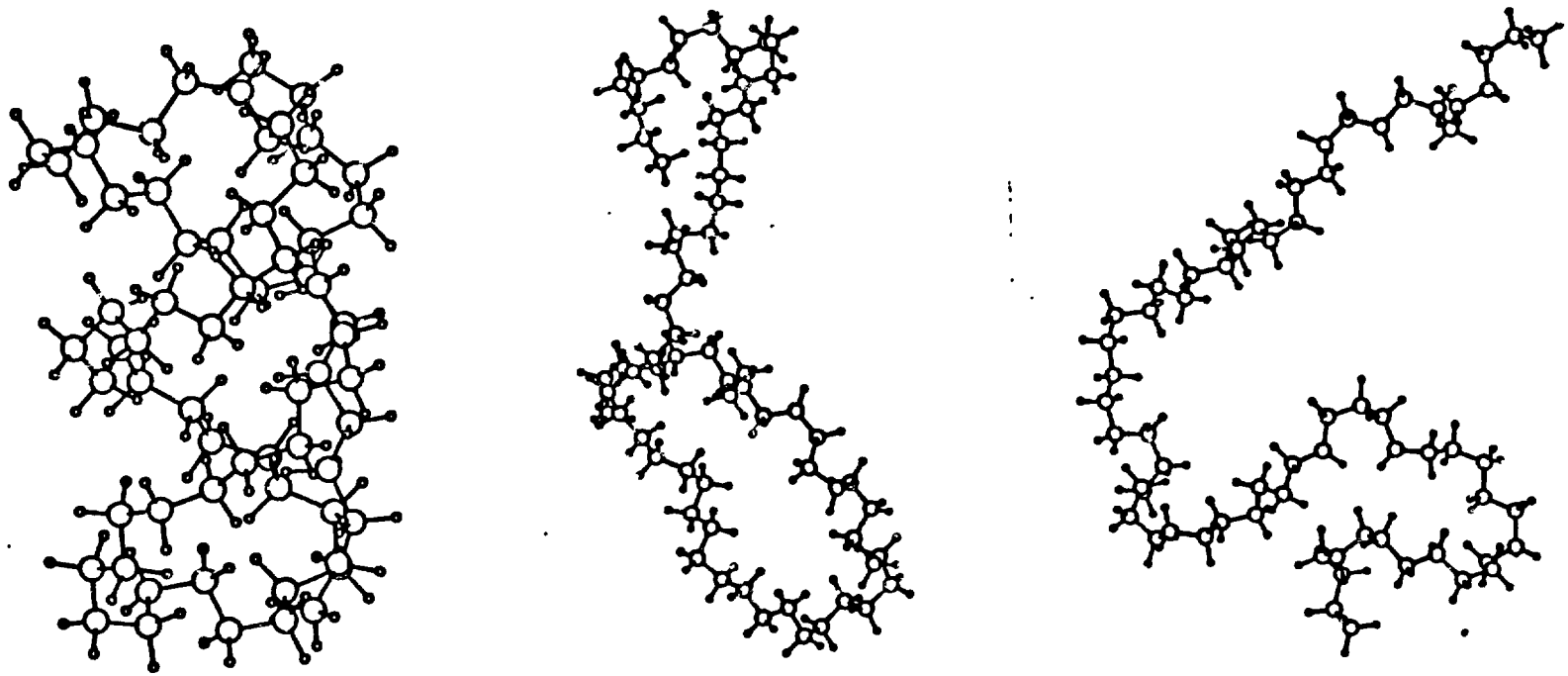


Figure 2. Snapshots of the conformations of $C_{64}H_{130}$ (polymethylene) at high temperature [Fig. 2(a)], where the system nearly behaves as a random coil, and at a temperatures (300°) below the globularization transition.

temperature, θ. Thisis indicated in Figure 3. As these studies proceed, we will consider more complicated side chains and expect to be able to make reliable predictions on biopolymers in two years.

D. Macroscopic Simulations

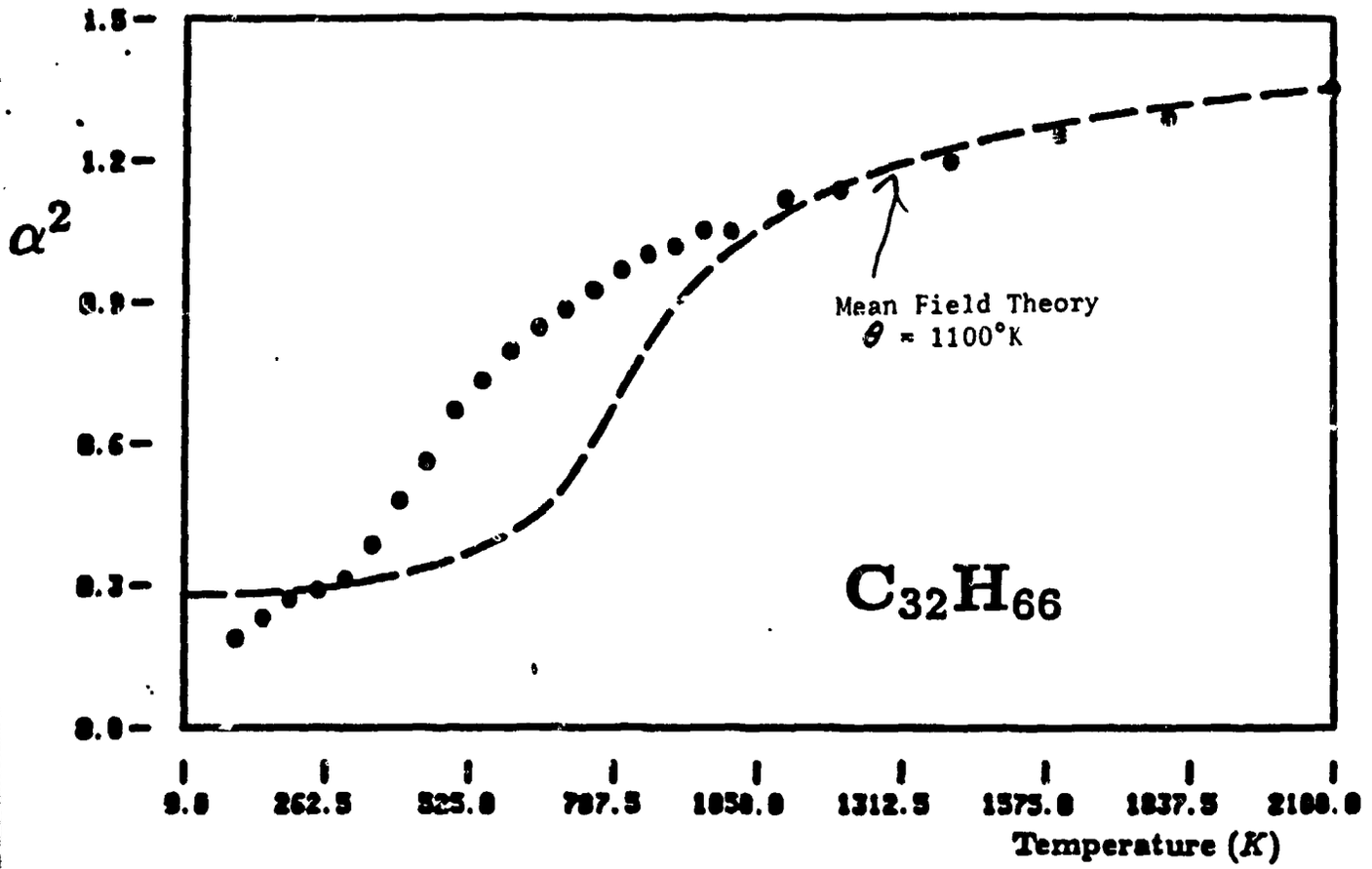
The microscopic atomic level views described above are essential for predicting the binding at specific sites; however, many considerations of DNA involve much more macroscopic concepts. For example, studies on circular DNA have shown that inducing sufficient extra twists can cause the DNA to writhe into a figure 8 (or a telephone cord). Experiments providing information about the energetics are available for systems with 2000 base pairs (100,000) atoms. Similarly, considerations of DNA packaging in the nucleus involve a number of nucleosomes, each of which has 200 base pairs (counting the connecting pieces) or 10,000 atoms per nucleosome. For systems of the size 100,000, it is not practical or useful to consider every atom. Instead, our approach here is to treat the DNA as a solid bar that has major and minor grooves inscribed but where distortions are described in terms of elastic (or inelastic) constants for twisting, bending helicity, etc. The parameters for the macroscopic model are derived from the microscopic simulations.

To simplify the description of the conformational behavior of double-stranded DNA, our method uses a compact shorthand representation of the molecule's conformational state. Here the molecule is divided into short length elements whose individual conformational states are specified by the local bending, helicity, and axial twisting rates. The conformational state of the entire molecule is then specified by the set of these conformational parameters for all the length elements. The conformational behavior of the DNA macromolecule is determined by the energy function over its conformational space

The evaluation of this energy function can be vastly simplified for a conformation specified in the shorthand representation compared with the usual representation giving explicit specification of the locations of all the atoms. The energy calculation is based on the full atomic-level energy expression, which is evaluated and tabulated for many conformations of short DNA segments (ten base pairs) and then described in terms of polynomials as a function of bending, torsion, and helicity. Evaluation of the energy of a macromolecular conformation is then accomplished by evaluating the energy polynomial for each of its length elements and summing.

E. Facilities

During the first year, the techniques for predicting tertiary structure have moved much faster than expected. The computer systems with parallel processing capabilities has provided the required computer speed. The new Evans & Sutherland PS 390 Graphics Systems in the Computer Simulation Laboratory and the Evans & Sutherland PS 330 in the Organic Chemistry Laboratory have provided the ability for all researchers to get involved in modeling. The modeling and simulations are providing a focus for the HIN-recombinase and single-stranded DNA binding protein projects. We are poised at a point where there should be trememdous progress over the next year.



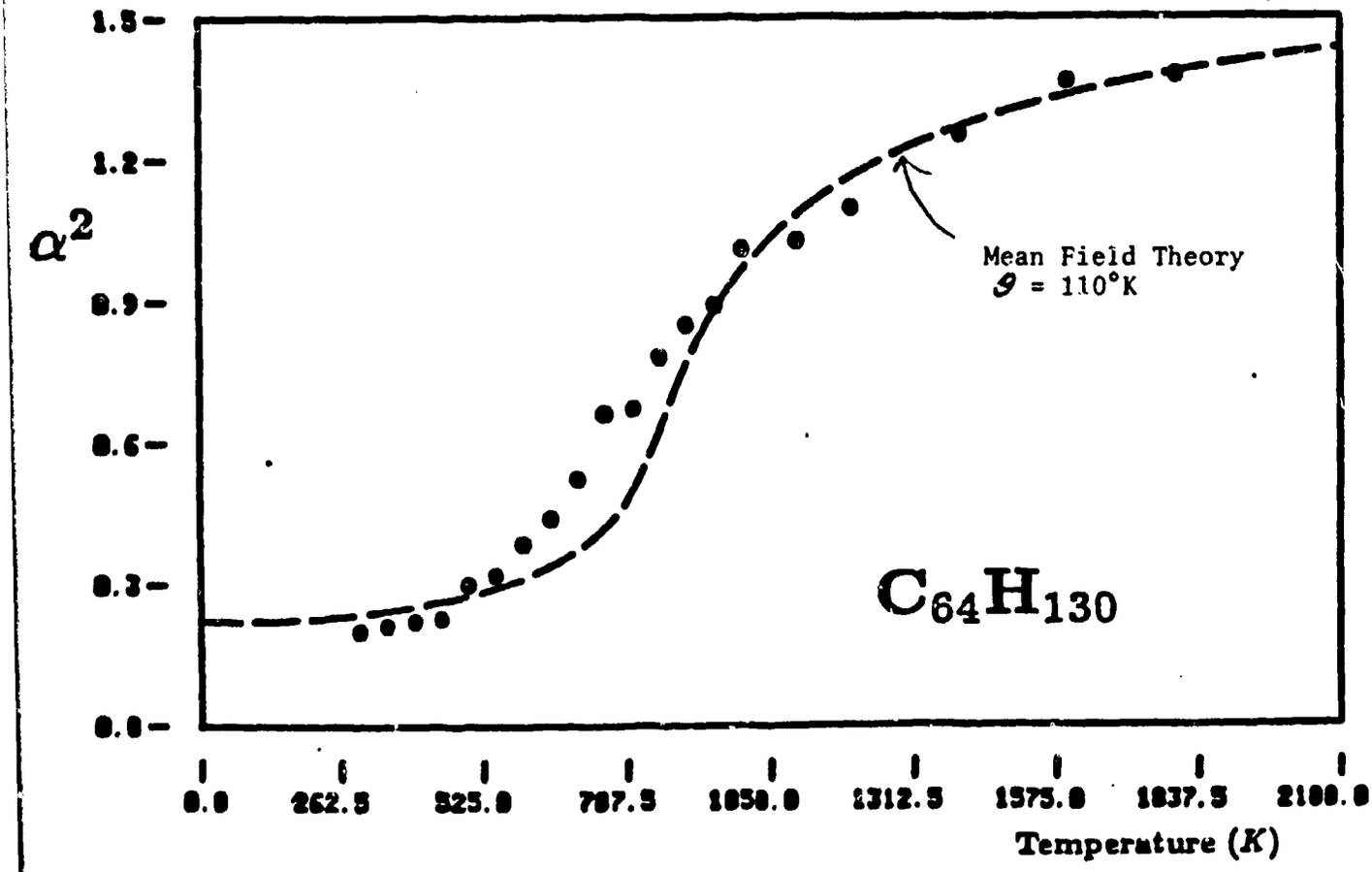


Figure 3. Relative size of $C_{64}H_{130}$ as a function of temperature. α^2 is the ratio of the calculated radius of gyration to that expected for a random coil. The dotted line shows how well the simulations are fit in terms of a one-parameter mean field theory.

F. Publications

1. Karasawa, N. and Goddard III, W. A. Phase transitions in polymethylene single chains (from Monte Carlo Simulations). *J Phys. Chem.*, to be submitted.

2. Karasawa, N. and Goddard III, W. A. Simulations of polypropylene conformations: dependence of properties atacticity. Manuscript in Preparation.

3. Mathiowetz, A. M. and Goddard III, W. A. Strategies for predictions of polypeptide folding: prospects for tertiary structure prediction. Manuscript in Preparation.

4. Mathiowetz, A. M. and Goddard III, W. A. Theoretical predictions of three-dimensional structures of cytochrome P-450 based on C_α-constrained torsionnal subspace optimization strategies.Manuscript in Preparation.

5. Mathiowetz, A. M. and Goddard III, W. A. Prediction of three-dimensional structures for rat and rabbit metallothioneins based on NMR-COSY and NOE data plus molecular dynamics/Monte Carlo simulations. Manuscript in Preparation.

Judith L. Campbell

Eukaryotes contain multiple species of proteins that bind preferentially but without sequence specificity to single-stranded DNA. Since the single-stranded DNA binding proteins (SSBs)[1] specifically stimulate certain DNA polymerases, it was at first thought that they were involved in unwinding DNA during replication, recombination and repair, as has been shown for their prokaryotic counterparts (for review see 1,2). Still, one protein carries out all these functions in bacteria and there was no good explanation for the multiplicity of SSBs in the eukaryotic systems. It is now clear that analogy to the prokaryotic prototype was much too simple a view and that the physical heterogeneity of eukaryotic SSBs reflects proteins of many diverse physiological functions. In fact, even in their mechanism of DNA polymerase stimulation, the eukaryotic and prokaryotic SSBs are very different. Prokaryotic SSBs stimulate cognate polymerases by specific protein-protein interactions. In eukaryotes, however, the stimulation apparently occurs because of a unique feature of DNA polymerase alpha, which unlike prokaryotic DNA polymerases and the other eukaryotic polymerases, is inhibited by naked single-stranded DNA. DNA polymerase α activity is not inhibited, however, if the single-stranded DNA is coated with an SSB (1).

By comparison of protein and DNA sequence data, the structural and functional relationships between the various eukaryotic SSBs that have been described over the years is just now coming to be understood. Interestingly, the ssDNA binding proteins described to date may not be SSBs at all. A few of these so-called SSBs have turned out to be dehydrogenases that fortuitously bind to DNA, presumably by virtue of their nucleotide binding sites (1). Furthermore, the importance of other SSBs in eukaryotic RNA metabolism has become obvious. Analysis of amino acid sequences of the prototype eukaryotic SSBs, UP1 and UP2 proteins from calf thymus (3) and HDP1 protein from mouse hepatoma (4) shows that there is extensive sequence homology with hnRNP proteins. Strongly homologous stretches of sequence can be seen between the hnRNP A1 protein and UP1 (5-7, 4), between the hnRNP A2 protein and HDP-1 (5), and UP2 is closely

related to yet another, non-A type, hnRNP protein (8). These hnRNP proteins are core components of the 40S hnRNP complex which contains nascent mRNA and is thought to be involved in transport or stabilization or processing of mRNA (5). Whether the same or related proteins are involved in such different processes as replication and RNA processing remains to be demonstrated.

We have shown that yeast, *Saccharomyces cerevisiae*, like other eukaryotes, contains multiple species of SSBs (9,10). To date, we have characterized a 45 kDa SSB, designated SSB1, most extensively. SSB1 was isolated on the basis of preferential binding to single-stranded versus double-stranded DNA and was subsequently shown also to bind to RNA (9). Although SSB1 seems to bind without sequence specificity, it is interesting to note that it copurifies through several affinity steps with the yeast poly (A) binding protein (11, Sachs and Jong, unpublished observations). SSB1 stimulates yeast DNA polymerase I on single-stranded DNA templates, which the polymerase by itself copies inefficiently (9). Because of this property, we felt that the protein might be involved in DNA replication or repair. In order to investigate this, we cloned the gene and carried out gene disruption experiments to see if the gene were essential and present in a single copy. Surprisingly, strains containing the gene disruption grow normally, even though they have been shown to contain no immunologically cross-reacting proteins. Furthermore, the gene is not required for sporulation, spore germination or recombination. DNA repair has not been tested. Because the abundance of the protein (20,000 copies per cell) suggests that it has an important biological role and because of the newly appreciated role of many eukaryotic SSBs in RNA metabolism we sought additional approaches to help elucidate the role of SSB1. In order to better define the function of yeast SSB1 the nucleotide sequence of the *SSB1* gene has been determined and compared to other proteins of known function. The amino acid sequence contains 293 amino acid residues, with M_r 32,853. There are several stretches of sequence characteristic of other eukaryotic single-stranded nucleic acid binding proteins. At the amino terminus, residues 39 - 54 are highly homologous to a peptide in calf thymus UP1, UP2 and a human hnRNP protein. Residues 125-162 comprise a 5-fold tandem repeat of the sequence RGGFRG, the composition of which suggests a nucleic acid binding site. Near to the C-terminus, residues 233 -245 are homologous to several RNA binding proteins. Ten out of eighteen C-terminal residues are acidic, a characteristic of the prokarytic SSBs and eukaryotic DNA and RNA binding proteins. In addition, examination of the subcellular distribution of SSB1 by immune fluorescence microscopy indicates that SSB1 is a nuclear protein, predominantly located in the nucleolus. Sequence homologies and the nucleolar localization make it likely that SSB1 functions in RNA metabolism *in vivo*, though an additional role in DNA metabolism cannot be excluded.

We have collaborated with John Abelson to extend these studies. The aromatic and basic amino acids in the RGGFRG repeat may be involved in nucleic acid binding. To test this, a peptide with the sequence,

RGGFRGGYRGGFRGRGRGNFRG

(base pairs 421-480), has been synthesized and its binding to DNA has been studied. A computer model is being formulated to predict possible interactions of the peptide with A-form DNA (see figures).

Finally, we have used SSB1 antibodies to probe the structure of the yeast nucleus. By immune fluorescent and nucleolar specific silver staining techniques we have determined that the single-strand DNA binding protein, SSB1, is located in the

yeast nucleolus. The amino acid sequence of SSB1 shares regions of homology with the mammalian nucleolins, nucleolar proteins that in turn share homology with hnRNPs and poly A binding proteins. We have also identified another single-strand DNA binding protein, SSB-36Kd, which binds silver under the conditions for nucleolar specific staining and is presumably another nucleolar protein.

We have used these nucleolar proteins as markers to examine the morphology of the nucleolus in one of the yeast RNA processing mutants, *rna*1. At non-permissive temperature, by SSB1 antibody fluorescent and nucleolar specific silver staining, the *rna*1 nucleolus appears disrupted. The staining patterns are no longer confined to a cap-like nucleolus filling one-third of the nucleus, as they are in the wild-type cells, but appear to be located over the entire nucleus. At the same temperature, wild-type and RNA splicing mutant, *rna*2, cells do not show this nucleolar disruption.

The *rna*1 mutation causes the accumulation of precursor rRNAs, precursor tRNAs and some precursor mRNAs. From these biochemical data, two models for the function of the *RNA*1 gene product have been proposed. In one model, the *RNA*1 protein is directly involved in the transport of mature RNAs out of the nucleus into the cytoplasm. In the other model, the *RNA*1 protein is somehow involved in a nuclear organization mechanism. The *rna*1 mutation then prevents proper alignment of the RNA processing enzymes and prevents them from functioning. Our data of a nucleolar disruption caused by the *rna*1 mutation supports the latter model of a nuclear organization mechanism involvement for the *rna*1 protein.

Footnote

[1]The term SSB has traditionally been used to describe proteins that bind ssDNA and whose *in vivo* role involves binding to DNA. This definition will be adhered to in this paper.

References

1. Chase, J. W. and Williams, K. R. (1986) Single-stranded DNA binding proteins required for DNA replication. *Ann. Rev. Biochem.* 55, 103-136.

2. Kilmartin, J.V. and Adams, A.E.M. (1984) Structural rearrangements of tubulin and actin during the cell cycle of the yeast *Saccharomyces*. *J. Cell Biol.* 98, 922-933.

3. Herrick, G., and Alberts, B. (1976) Purification and physical characterization of nucleic acid helix-unwinding proteins from calf thymus. *J. Biol. Chem.* 251, 2124-2132. Nucleic acid helix-coil transitions mediated by helix-unwinding proteins from calf thymus. *J. Biol. Chem.* 251 2133-2141. Herrick, G., Delius, H., and Alberts. (1976) Single-stranded DNA structure and DNA polymerase activity in the presence of nucleic acid helix-unwinding proteins from calf thymus. *J. Biol. Chem.* 251, 2142-2146.

4. Williams, K. R., Stone, K. L., and Lopresti, M. B., Merrill, G. M. and Planck, S. R. (1985) Amino acid sequence of the UP1 calf thymus helix-destabilizing protein and its homology to an analogous protein from mouse myeloma. *Proc. Natl. Acad. Sci. USA* 82, 5666-5670.

5. Kumar, A., Williams, K.R., and Szer, W. (1986) Purification and domain structure of core hnRNP proteins A1 and A2 and their relationship to single-stranded DNA-binding proteins. *J. Biol. Chem.* 261, 11266-11273.

6. Merrill, B. M., LoPresti, M. B., Stone, K. L., and Williams, K. R. (1986) High pressure liquid chromatography purification of UP1 and UP2, two related single-stranded nucleic acid-binding proteins from calf thymus. *J. Biol. Chem.* 261, 878-883.

7. Pandolfo, M., Valentini, O., Diamonti, G., and Riva, S. (1985) Single stranded DNA binding proteins derive from hnRNP proteins by proteolysis in mammalian cells. *Nucl. Acids Res.* 13, 6577-6590.

8. Lahiri, D., and Thomas, J. O. (1986) A cDNA clone of the hnRNP C protein and its homology with the single-stranded DNA binding protein UP2. *Nucl. Acids Res.* 14, 4077-4094.

9. Jong, A. Y., Aebersold, R., and Campbell, J. L. (1985) Multiple species of single-stranded nucleic acid-binding proteins in *Saccharomyces cerevisiae*. *J. Biol. Chem.* 260, 16367-16374.

10. Jong, A. S., and Campbell, J. L. (1986) Isolation of the gene encoding yeast single-stranded nucleic acid binding protein 1. *Proc. Natl. Acad. Sci. USA* 83, 877-881.

11. Sachs, A. B., Bond, M. W., and Kornberg, R. D. (1986) A single gene from yeast for both nuclear and cytoplasmic polyadenylate-binding proteins: domain structure and expression. *Cell* 45, 827-835.

John N. Abelson

We have made good progress in the characterization of the yeast tRNA ligase. The following reports detail this year's progress and present some of the approaches we will take in the future. In addition to the mapping of the domains, we have committed a major effort to increasing the scale of the purification. We now believe that we are in a position to purify 100 milligrams amounts of ligase and are embarked on such a purification. The eventual objective of this protein crystallization is obtaining a crystal structure.

A. Yeast tRNA Ligase Sequence: Studying the Domains

tRNA ligase is one of two enzymes required for tRNA splicing in *Saccharomyces cerevisiae*. The enzyme is likely a single polypeptide of 95 kDa, and possesses three activities responsible for the second stage of the tRNA splicing mechanism: a cyclic phosphodiesterase; a polynucleotide kinase; and a ligase activity. The tRNA ligase protein has been purified and its gene has been cloned (Phizicky *et al.*, 1986). We have sequenced the gene and analyzed the transcripts from this region (Westaway, Phizicky, and Abelson, manuscript in preparation). The inferred amino acid sequence of the tRNA ligase reading frame reveals a basic protein of 827 amino acids, which at present is not homologous to other known proteins of similar function. We have found two other open reading frames, ORF1 and ORF2, near the 5' end of the tRNA ligase gene.

We are pursuing the study of the tRNA ligase protein by attempting to dissect the domains responsible for the four splicing activities listed above. Fragments of the tRNA ligase protein have been isolated, each of which contains only some of the different activities of the intact polypeptide, but all of which are recognized by anti-tRNA ligase antibody. The amino-terminal amino acid sequence for three of these purified fragments (65 kDa, 40 kDa, and 25 kDa) has been obtained. We are cloning these fragments using the DNA sequence from the ligase gene, by inserting appropriate stop and start codons using oligonucleotide-directed mutagenesis and removing all other DNA sequences of the mature tRNA ligase protein. By overproducing these peptide fragments in *E. coli*. we remove any possibility of contamination by other portions of the tRNA ligase protein or other yeast proteins. Since the fragments resulting from intact tRNA ligase are recognized by anti-tRNA ligase antibody, we should be able to detect the cloned fragments of the protein as well. Purification of these cloned peptide fragments, followed by assays for each separate enzymatic activity, should verify the result obtained using the peptide fragments of intact tRNA ligase. In this manner we hope to effect the localization of the adenylation site of tRNA ligase, as well as a systematic elucidation of the distinct domains of tRNA ligase necessary for each function during the entire splicing reaction.

Reference

Phizicky, E. M., Schwartz, R. C. and Abelson, J. (1986) *J. Biol. Chem.* **261**, 2978-2986.

B. Activity Domains of Yeast tRNA Ligase

The splicing reaction of yeast tRNA proceeds in two distinct stages. In the first stage, there are two precise endonucleolytic cleavages at the mature sequence boundaries to produce the tRNA halves and linear intervening sequence. In the second stage of the splicing reaction, the tRNA halves are joined in an ATP-dependent step. All four of the enzymatic activities required in the second stage of splicing reaction--an adenylated enzyme intermediate, a tRNA ligase, a polynucleotide kinase and a cyclic phosphodiesterase-cosediment on a glycerol gradient with a single 90 kDa polypeptide.

The *Saccharomyces cerevisiae* tRNA ligase has been cloned and expressed in *E. coli* on a plasmid containing the ligase gene under the control of the *tac* inducible promoter. This ligase expressed in *E. coli* is a fully active, 90 kDa polypeptide. Along with the 90 kDa polypeptide there are a number of discrete lower molecular weight fragments that are antigenically related to tRNA ligase.

These major fragments can be produced by trypsin digestion of the 90 kDa ligase protein. We have purified the three major fragments (65, 40 and 25 kDa molecular weight) to homogeneity. Amino acid sequences of the 65 kDa and 40 kDa fragments reveal these two fragments to have the identical amino terminal sequence as the 90 kDa ligase protein, but 25 kDa protein is from the C-terminus of the protein The 90 kDa ligase as well as the 65 kDa and 40 kDa fragments can be adenylated while the 25 kDa fragment cannot. The 65 kDa fragment possesses the majority of the polynucleotide kinase activity, whereas the 25 kDa fragment has less than 5% of the original kinase ability, and the 40 kDa fragment has no kinase activity. The 65, 40 25 kDa fragments do not have the ability to catalyze the ligation of tRNA-half molecules. Addition of the 25 kDa fragment with the 65 kDa fragment restores tRNA ligase activity. These data indicate that yeast tRNA ligase has

distinct activity domains that must act in concert to achieve the complete tRNA ligation reaction.

Leroy E. Hood

The peptide synthesis portion of this grant encompasses two roles: to improve methods for the total chemical synthesis of the small DNA-binding proteins/large peptides to be designed and studied; and, to pass that technology on to other members of the collaboration.

A. Background

Over the years 1977-1983, we developed an improved understanding of the basic principles of stepwise solid phase peptide synthesis. This included developing mechanistic insight into the series of chronic chemical side reactions affecting every step of the synthesis and devising new chemistries to minimize their occurrence (1-3). In this way, the level of side reactions was reduced from an average of 4% per amino acid residue to 0.1% per residue (4).

Beginning in 1980, we also developed novel insights into the phenomenon of "difficult" sequences in stepwise SPPS. This was possible because the obscuring background of chemical side reactions had been eliminated. Between 1980 and 1983 we developed an empirical description of the phenomenon and a mechanistic understanding of its molecular origins that enabled us to adopt new chemical strategies that substantially reduced the severity of the problem, and for the first time made available general synthetic chemistry for the synthesis of any amino acid sequence (5-7).

In 1983/84 this work led to the design and development the first truly automated peptide/protein synthesizer based on these mechanistic insights and improved chemistries, in collaboration with scientists and engineers at Applied Biosystems (8, 9).

Between Fall 1984 and Fall 1986, we developed highly optimized chemistry for use on this instrument (10, 11), and applied this to the synthesis and structure/function studies of a number of biologically active peptides and proteins (12-14). This chemistry had considerable generality and was quite efficient, averaging 99.4% yield per amino acid residue in the synthesis of polypeptide chains 50-140 residues in length.

B. Progress Report

1. Technology Transfer

We have trained David Mack, a graduate student in Peter Dervan's laboratory, in the highly optimized chemistry described above. First, in the optimized manual synthetic protocols and in our methods for the deprotection and work up of the peptide products. He and others in the Dervan laboratory have used these protocols for the synthesis of DNA binding proteins under this grant. We have also trained David Mack in the use of the automated synthesizer and our automated chemistry protocols. He has used this to carry out several syntheses of DNA binding domains and to evaluate the performance of the chemistry and the instrument. It is expected that this knowledge will also be transferred to the Dervan group for their own use.

2. Improved Chemistry

Even the highly optimized chemistry described above has a number of shortcomings. The chemistry was slow, requiring 1.5-2 hours per residue for an average of 15 residues per day even with the continuous 24 hour per day unattended operation possible with the automated synthesizer. In addition, the chemistry was expensive at $15-$25 for consumables per residue. Both the cost and the lack of speed, together with the substantial purification problem arising from the yield of 99.4% per residue, were substantial disadvantages for the chemical synthesis of proteins and their analogs. In the past grant year, we have addressed these problems as follows.

Our chemistry for automated stepwise SPPS was reworked in the light of our current state of knowledge. This resulted in a number of fundamental changes in the synthetic protocols. The numerous batchwise washing operations involved in standard SPPS were replaced with single, short "flow washes," resulting in higher efficiency and more rapid synthesis. In a successful attempt to speed up the synthetic cycle, 100% TFA was used for the N-alpha Boc removal step and very high concentrations of Boc-amino acid symmetric anhydrides were used in the coupling step. The peptide-resin came in contact with only a single, highly efficient solvent (DMF) throughout the synthetic cycle.

These changes increased the efficiency of chain assembly by several tenths of a percent per residue. The improved protocols were fully compatible with the existing automated synthesizer and enabled us to reduce the cycle time to 25 minutes per amino acid residue, and to use only single couplings for peptides up to about 50 residues in length (15). We also developed a reduced scale of synthesis for the automated instrument, at 20 minutes per residue, that allowed the rapid synthesis of several hundred milligrams of crude peptide or up to 50 milligrams of purified peptide, at a cost of $3-$5 per residue. This has obvious utility for pilot syntheses and for the rapid production of analogs (16).

3. Current Work

We are adapting the principles of the new, more efficient rapid chemistry protocols to the synthesis of the long polypeptide chains (>50 residues) that form structural domains in proteins. This work is partly completed, and will result in protocols for the automated synthesis of long polypeptide chains that will be about twice as fast as those we currently use, i.e., about 45-60 minutes per residue, for approximately 30 residues per day. We anticipate that these will also give higher yields of chain assembly as was the case for the peptide synthesis protocols.

We are also exploring strategies for "capping" during chain assembly. This will eliminate deletion peptide formation, and in combination with a novel strategy for selecting only full length peptide chains, will allow rapid purification of the target polypeptide chains. This work will take advantage of our large data base of peptide syntheses, estimated at over 320,000 documented couplings in the past 3 years. Of particular importance will be the use of calibrated crude product analyses for a number of peptides to check that the capping is not introducing side reactions and is eliminating deletion peptides, and that the purification strategy is actually removing the terminated peptides from the product mixture.

It is anticipated that these improved synthetic methods will extend our ability to prepare high purity synthetic proteins from the current limit of about 50 residues

(e.g., TGF-alpha, see attached manuscript) to the 150 residue range. Such high purity products are essential for many of the studies proposed under this grant. After careful checking, the protocols will be evaluated in protein synthesis and then made available to our collaborators under this grant.

References

1. Kent, S. B. H., Mitchell, A. R., Engelhard, M. and Merrifield, R. B. (1979) *Proc. Natl. Acad. Sci. USA* 76, 2180-2184.

2. Kent, S. B. H. and Merrifield, R. B. (1983) *Int. J. Pep. Prot. Res.* 22, 57-65.

3. Kent, S. B. H. (1984) in *Peptides: Structure and Function*, Proceedings of the Eighth American Peptide Symposium (Hruby, V. J. and Rich, P. H., Eds.), pp. 99-102, Pierce Chem. Co., Rockford, IL.

4. Kent, S. and Clark-Lewis, I. (1985) in *Synthetic Peptides in Biology and Medicine* (Alitalo, K. Partanen, P. and Vaheri, A., Eds.), pp. 29-57, Elsevier, Amsterdam.

5. Kent, S. B. H. and Merrifield, R. B. (1981) in *Peptides 1980: Proceedings of the 16th European Peptide Symposium* (Brundfelt, K. Ed.), pp. 328-333, Scriptor, Copenhagen.

6. Meister, S. M. and Kent, S. B. H. (1984) in *Peptides: Structure and Function*, Proceedings of the Eighth American Peptide Symposium (Hruby, V. J. and Rich, D. H. Eds.), pp. 103-106.

7. Kent, S. B. H. (1985) in *Peptides: Structure and Function*, Proceedings of the Ninth American Peptide Symposium (Deber, C. M., Hurby, V. J. and Kopple, K. D., Eds.), pp. 407-414, Pierce Chem. Co., Rockford, IL.

8. Kent, S. B. H., Hood, L. E., Beilan, H., Meister, S. and Geiser, T. (1984) in *Peptides 1984: Proceedings of the 18th European Peptide Symposium* (Ragnarsson, U., Ed.), p. 185-188, Alqvist and Wiksell, Stockholm.

9. Kent, S. B. H., Hood, L. E., Beilan, H., Bridgeham, J., Marriot, M., Meister, S. and Geiser, T. (1985) in *Peptide Chemistry 1984: Proceedings of the 22nd Japanese Peptide Symposium* (Isumiya, N., Ed.), pp. 217-222, Protein Research Foundation, Osaka.

10. Clark-Lewis, F., Aeberwsold, R. A., Ziltner, H., Schrader, J. W., Hood, L. E. and Kent, S. B. H. (1986) *Science* 231, 133-139.

11. Clark-Lewis, F. and Kent, S. B. H. (1987) in *Receptor Biochemistry and Methodology: The Use of HPLC in Protein Purification and Characterization* (Kerlavage, A. R., Ed.), in press.

12. Kent, S. B. H. and Parker, K. F. (1987) in *Therapeutic Peptides and Proteins: Assessing the New Technologies*, Proceedings of the Banbury Center Conference at Cold Spring Harbor, in press.

13. Neurath, A. R., Kent, S. B. H., Strick, N. and Parker, K. (1986) *Cell* 46, 429-436.

14. Clark-Lewis, F., Hood, L. E. and Kent, S. B. H. Submitted for publication.

15. Kent, S. B. H., Parker, K. F., Schiller, D. L., Woo, D. D.-L., Clark-Lewis, K. and Chair, B. T. (1987) Proceedings of the 10th American Peptide Symposium (Marshall, G. R., Ed.), in press.

16. Kent, S. B. H. and Parker, K. F. Manuscript in preparation.